CS 7545: Machine Learning Theory **Fall 2021**

Lecture 1

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Statistical Inference

Suppose $x_1, x_2, \ldots, x_m \in \mathbb{R}$. Then we can use the following statistics:

- Mean: $\mu = \frac{1}{m} \sum x_i$,
- Variance: $\sigma^2 = \frac{1}{m} \sum (x_i - \mu)^2$.

In high dimension, say $\mathbb{R}^d$, we must be precise about variance. We will project each point onto a unit vector $\boldsymbol{v} \in \mathbb{R}^d$:

$$\sigma_{\boldsymbol{v}}^2 = \frac{1}{m} \sum_i (\boldsymbol{x}_i \cdot \boldsymbol{v} - \boldsymbol{\mu} \cdot \boldsymbol{v})^2 \tag{1}$$

$$= \frac{1}{m} \sum_i ((\boldsymbol{x}_i - \boldsymbol{\mu}) \cdot \boldsymbol{v})^2 \tag{2}$$

$$= \boldsymbol{v}^\intercal \left(\frac{1}{m} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\intercal \right) \boldsymbol{v}. \tag{3}$$

The middle term is $\boldsymbol{\Sigma}$, the covariance matrix. We can write

$$\sigma_{\boldsymbol{v}}^2 = \boldsymbol{v}^\intercal \boldsymbol{\Sigma} \boldsymbol{v}. \tag{4}$$

Of course, the mean and variance does not characterize the data distribution. One way to properly characterize the distribution is by estimating the probability density function (PDF). For example, the uniform distribution in $[a, b]$ has density

$$p(x) = \begin{cases} \frac{1}{|b-a|} & x \in [a, b], \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

By definition, a PDF is a nonnegative integrable function which integrates to 1 over the domain.

2 The Gaussian Distribution

2.1 The Standard Gaussian

The one-dimensional standard Gaussian $\mathcal{N}(0,1)$ has PDF

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{6}$$

In $\mathbb{R}^d$, the distribution is called the multivariate standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The PDF is the product of 1-dimensional Gaussians along each coordinate:

$$p(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^d} e^{-\|\boldsymbol{x}\|^2/2} \tag{7}$$

$$= \prod_i \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2}. \tag{8}$$

Fact 2.1: Spherical Symmetry of the Multivariate Standard Gaussian

For $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, we have $\boldsymbol{x} \cdot \boldsymbol{v} \sim \mathcal{N}(0,1)$. Equivalently,

$$\Pr[\boldsymbol{x} \cdot \boldsymbol{v} = t] = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{9}$$

Proof: We have

$$\Pr[\boldsymbol{x} \cdot \boldsymbol{v} = t] = \int_{\boldsymbol{x}:\boldsymbol{x}\cdot\boldsymbol{v}=t} \frac{1}{(\sqrt{2\pi})^d} e^{-\|\boldsymbol{x}\|^2/2} dx. \tag{10}$$

We can write $\boldsymbol{x}$ in the basis $(\boldsymbol{v}, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d)$ so that

$$\|\boldsymbol{x}\|^2 = (\boldsymbol{x} \cdot \boldsymbol{v})^2 + \|\boldsymbol{y}\|^2. \tag{11}$$

Then,

$$\Pr[\boldsymbol{x} \cdot \boldsymbol{v} = t] = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_{\boldsymbol{y} \in \mathbb{R}^{d-1}} \frac{1}{(\sqrt{2\pi})^{d-1}} e^{-\|\boldsymbol{y}\|^2/2} dy. \tag{12}$$

Since the integrand is a $(d-1)$-dimensional Gaussian, its integral over $\mathbb{R}^{d-1}$ is 1 by definition. We have to be careful about the differential here by checking the determinant of its Jacobian, but since $(\boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_d)$ is orthogonal, we immediately have $dy = dx$. ■

Question 2.1

Where does the $\frac{1}{\sqrt{2\pi}}$ come from in the denominator of the PDF?

Fact 2.2: Gaussian PDF Integrates to 1

Let $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then, $\int_{\mathbb{R}} p(x)dx = 1$.

Proof: This is a difficult integral *a priori*, so we consider the two-dimensional case:

$$p(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} \tag{13}$$

$$\int_{\mathbb{R}^2} p(x, y)dxdy = \left(\int_{\mathbb{R}^2} p(x)dx \right)^2. \tag{14}$$

We take advantage of spherical symmetry by integrating in spherical coordinates (in two dimensions, these are concentric circles).

$$\left(\int_{\mathbb{R}^2} p(x)dx \right)^2 = \frac{1}{2\pi} \int_{r=0}^{\infty} 2\pi r e^{-r^2/2} dr \tag{15}$$

$$= \int_{r=0}^{\infty} r e^{-r^2/2} dr. \tag{16}$$

Let $u = r^2/2$, then

$$\int_{r=0}^{\infty} r e^{-r^2/2} = \int_{u=0}^{\infty} e^{-u} du \tag{17}$$

$$= 1. \tag{18}$$

■

2.2 The General Gaussian

The one-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2)$ has PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right). \tag{19}$$

The multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has PDF

$$p(x) = \frac{1}{(\sqrt{2\pi})^d |\det(\boldsymbol{\Sigma})|} \exp\left(\frac{-(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2} \right), \tag{20}$$

with covariance matrix $\boldsymbol{\Sigma}$ positive definite (since we cannot have negative variance).

In a standard Gaussian, the densities are symmetric over the sphere. In a general Gaussian, they are symmetric over an ellipsoid. That is, a general Gaussian is a standard Gaussian acted on by a linear transformation described by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Fact 2.3: Linear Transformations of a Gaussian

Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then for $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$, $\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{b}, \mathbf{A}\mathbf{A}^\intercal)$. In the other direction, let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

3 Maximum Likelihood Estimation for Gaussians

Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \in \mathbb{R}^d$ and we would like to find the best-fit Gaussian. In the case when $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$, we would like to find $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ which maximizes the likelihood of generating the observed data. The log-likelihood function we will maximize is

$$L(\boldsymbol{\mu}, \sigma^2) = \log \prod_{j=1}^{m} \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right) \tag{21}$$

$$= -\Big(md \log(\sqrt{2\pi}) + md \log(\sigma) + \frac{1}{2\sigma^2} \sum_j \|\boldsymbol{x}_j - \boldsymbol{\mu}\|^2\Big). \tag{22}$$

Lemma 3.1: Maximum Likelihood Estimation for Gaussians

The sample mean $\hat{\boldsymbol{\mu}}$ and sample variance $\hat{\sigma^2}$ are maximum likelihood estimators for the population mean $\boldsymbol{\mu}$ and population variance σ^2.

Proof: For the variance,

$$\frac{\partial L}{\partial \sigma} = -\frac{md}{\sigma} + \frac{1}{\sigma^3} \sum_j \|\boldsymbol{x}_j - \boldsymbol{\mu}\|^2. \tag{23}$$

Setting equal to zero implies

$$\sigma^2 = \frac{1}{md} \sum_j \|\boldsymbol{x}_j - \boldsymbol{\mu}\|^2 = \hat{\sigma^2}. \tag{24}$$

For the mean,

$$\frac{\partial L}{\partial \mu_i} = -\frac{1}{2\sigma^2}\Big(-2 \sum_j (x_{ji} - \mu_i)\Big). \tag{25}$$

Setting equal to zero implies

$$\mu_i = \frac{1}{m} \sum_j x_{ji} = \hat{\mu_i}. \tag{26}$$

■

4 The Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental result which states that the distribution of sums (or equivalently, means) of independent random variables tends towards a Gaussian. There

are many quantitative formulations of the CLT, but the following provides a useful bound based on the first three moments.

Theorem 4.1: Berry-Esseen CLT

Let $X_1, X_2, \ldots, X_n$ be independent random variables. Let $Y_n = \sum_{i=1}^n X_i$ and $Z_n \sim \mathcal{N}(\mathbb{E}[Y_n], \mathrm{Var}(Y_n))$. Then for $t \in \mathbb{R}$ and $c \in [0, 1)$,

$$\left| \Pr[Y_n \leq t] - \Pr[Z_n \leq t] \right| \leq \frac{c}{\mathrm{Var}(Y_n)^{3/2}} \sum_i \mathbb{E}[|X_i|^3]. \tag{27}$$

Example. Suppose $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(-1, 1)$. Then $\mathbb{E}[|X_i|^3] = 1$ and $\mathrm{Var}(Y_n) = n$. By the theorem,

$$\left| \Pr[Y_n \leq t] - \Pr[Z_n \leq t] \right| \leq \frac{c}{n^{3/2}} n \tag{28}$$

$$= \frac{c}{\sqrt{n}}. \tag{29}$$